心理学报 2021, Vol. 53, No. 2, 155−169
Acta Psychologica Sinica

https://dx.doi.org/10.3724/SP.J.1041.2021.00155

小学生羞怯特质预测及语言风格模型构建*

骆 方[1] 姜力铭[1] 田雪涛[2] 肖梦格[1] 马彦珍[3] 张 生[3]

([1]北京师范大学心理学部, 北京 100875) ([2]北京交通大学计算机与信息技术学院, 北京 100044)
([3]中国基础教育质量监测协同创新中心, 北京 100875)

摘 要 收集在线教学平台上 1306 名小学生的作文、日记及评论, 采用自然语言处理技术进行文本分析, 并应用机器学习模型实现对羞怯特质的自动预测, 构建小学生羞怯行为、认知和情绪的语言风格模型。研究发现：(1)扩充的心理词典适合分析小学生文本; (2)分别存在羞怯行为、认知和情绪问题的学生其日常用语既有共性也有特性, 且与普通学生存在差异; (3)羞怯各维度在不同分类器上达到较好的预测效果, 其中随机森林模型的整体表现相对最好。

关键词 羞怯, 在线写作, 心理词典, 文本挖掘, 语言风格模型
分类号 B844

1 引言

羞怯是一种普遍存在的主观体验, 每个人在日常生活中都会不同程度地感受到羞怯。小学阶段是羞怯特质形成的关键时期, 持续且较高水平的羞怯对学生的社会交往、人格发展、身心健康等方面均具有消极影响。因此, 对早期羞怯的测量尤为重要。以往对羞怯的测量主要为自陈量表法和观察法, 二者均存在一定的问题：被试容易对自陈量表作假回答(李亚红 等, 2005; 骆方, 张厚粲, 2007), 量表法也不适用于长期追踪和动态反映学生羞怯的发展。观察法虽然能够评估个体是否具有羞怯行为, 但是难以评估羞怯的认知和情绪体验, 这些特征是内隐的, 不易被观察到(Leary, 1986)。

基于个体留下的网络痕迹来预测人格品质是近些年的研究热点(Azucar et al., 2018), 通过对个体在线数据的挖掘构建预测模型, 能够持续输出对被试的评估结果, 比如, 研究者基于被试在微博上发布的内容构建抑郁预测模型(白朔天 等, 2014), 从而及时甄别博主的自杀倾向并进行及时干预。羞怯与焦虑和抑郁一样具有较高的内隐性, 除非有明显的羞怯行为表现出来, 大部分内心有强烈的羞怯认知和情绪体验的个体难以被直接观察到, 但是他们的日常用语可能与普通个体存在差别。小学生的作文及日记中包含着其日常学习、生活的真实经历和感受, 羞怯相关事件及体验也能够被记录在内。因此, 本研究基于小学生的在线作文、日记和评论, 拟采用自然语言处理以及机器学习技术, 挖掘羞怯学生的语言风格, 即羞怯学生日常用语中的词汇使用习惯和特点, 构建羞怯的语言风格模型, 并据此建立羞怯特质的自动预测模型, 为实时监测学生的羞怯状况并进行干预提供可能。

1.1 羞怯及其测量

Lewinsky (1941)最早提出“羞怯”这一构念, 将其界定为一种心理状态：个体极度抑制, 感到自卑、被忽略, 对自己的感觉与情绪过分敏感, 同时伴随有许多常见的生理症状, 如脸红、出汗、发抖、说话不流利等; Zimbardo 等人(1975)提出, 羞怯是指个体不愿意接近他人, 或不愿意进入到被他人关注的、难以回避的情境中, 强调了“不愿意”的行为反应倾向; Cheek 和 Buss (1981)将羞怯界定为在人际交往的情境中, 个体由于感觉到他人在评价自己而

收稿日期: 2020-04-22
* 国家自然科学基金联合基金项目(U1911201)。
通信作者: 张生, E-mail: zhangsheng@bnu.edu.cn

产生紧张、害怕、尴尬或困窘的情绪。从上述定义可以看出，羞怯包含了行为、认知、情绪及生理反应多个维度的表现。

羞怯从儿童五、六岁左右开始发展(张昌 等, 2012)。儿童期羞怯对青春期的外倾性、开放性、情绪稳定性(Baardstu et al., 2019)、人际关系(Karevold et al., 2012)、内在问题(如焦虑、低自尊等) (Caspi et al., 1988)以及各领域的适应问题(Rubin et al., 1995)均具有预测作用。羞怯个体往往缺乏足够的应对策略，倾向于情绪化地应对问题，如攻击、自我伤害等(Eisenberg et al., 1998; Findlay et al., 2009; Kagan, 1997)，导致更加严重的社会问题。然而，早期羞怯的消极影响和测量尚未得到教育工作者的充分关注，这一方面是由于羞怯的学生沉默寡言，符合课堂规范(Evans, 2010)，通常成为老师眼中安静顺从的好学生，导致羞怯所带来的消极影响难以被发现；另一方面是由于中国传统文化鼓励约束和克制的态度或行为，羞怯的相关表现在某种程度上被认为是个体在社交方面的成熟(Chen, 2010; Ho, 1986)。因此，在我国文化背景下，对早期羞怯进行及时、有效的测量尤为重要。

研究者认为，羞怯的测量应当全面捕捉羞怯在行为、认知、情绪、生理四个维度上的表现(Henderson et al., 2014)，因为人们在羞怯情感和羞怯行为上的困扰程度往往不同(Pilkonis, 1977)，许多羞怯个体能够掩饰羞怯的内在焦虑体验，抑制羞怯的行为反应，从而维持自身的社会功能，这些个体被称为“外向羞怯者” (Zimbardo, 1977)。目前，羞怯的主流测量方法是自陈量表法。国外应用较广的羞怯量表有 Zimbardo (1977)编制的斯坦福羞怯量表(Stanford Shyness Scale)以及 Cheek 和 Buss (1981)编制的羞怯量表(Cheek & Buss Shyness Scale)。国内研究者多采用徐春蓉 (2001)编制的《国小儿童害羞量表》，例如，孙源泉等人(2009)采用该量表研究震区丧亲儿童的羞怯、创伤后应激障碍症状和心理健康之间的关系。自陈量表能够全面地捕捉羞怯在行为、认知、情绪、生理四个维度上的表现，但该方法难以实现对羞怯特质的重复测量和持续监测。除自陈量表法外，研究者还通过生理指标法、行为观察法等方式对羞怯进行测量，这两种方法均无法全面地测量羞怯各个维度：生理指标法(例如心率)仅能捕捉到羞怯的情绪及生理反应，且更适用于测量状态性羞怯而非特质性羞怯(Brodt & Zimbardo, 1981; Martin, 1961)；行为观察法仅能测量到羞怯的外显行为，难以测量内隐的羞怯情绪和认知，且个体表现出的羞怯行为可能受到个体的控制(Henderson et al., 2014)，因此，观察的结果不完全可靠。

为进一步验证行为观察法的实际有效性，检验教师的日常观察是否能够有效测量学生的羞怯水平，本研究选择华西小学三年级 2 班的 49 名学生作为预研究样本，要求该班学生作答《国小儿童害羞量表》，作为学生自评数据。同时，要求班主任从“害羞行为”、“害羞认知”、“害羞情绪反应”三个维度对该班学生进行 7 点评分，作为教师评定数据。结果显示，教师评分与学生自评之间相关较低。由此可见，班主任的观察难以有效地识别羞怯学生，因此，我们需要寻找更加有效的测评手段，捕捉学生内隐的羞怯认知和情绪。

语言是人们表达内在的思想和情绪的有效途径(Tausczik & Pennebaker, 2010)，小学生的作文及日记是学生在自然状态下的自我表达，文本的内容主要为日常学习、生活的真实经历，羞怯相关事件及体验也能够被充分记录在内。因此，本研究将探索一种全新的羞怯测量方法，即通过获取小学生在较长时间内的大量作文及日记文本，采用文本挖掘方法建立监测个体羞怯特质的模型。

1.2 文本挖掘预测心理特质

羞怯是一种典型的人格特质，人格与语言有密切的关系(Mairesse et al., 2007; Pennebaker & King, 1999)，尤其是与社交性相关的特质(Allport, 1937; Cattell, 1943; Goldberg, 1982; Norman, 1963)。对自然语言的分析能够揭示该语言背景下个体的语言特征与心理特质之间的关系，进而通过语言模式预测心理特质(Yarkoni, 2010)，而词汇特征则是个体的思维、感觉、观点和人格特质的重要线索(Argamon et al., 2005)。

以往基于自然语言的心理学研究受到文本数据收集的局限，通常要求被试在实验室中完成命题写作，例如，描述个人的过去经历或未来计划(Fast & Funder, 2008; Hirsh & Peterson, 2009)。然而，这种方式收集到的文本受情境、新近经验等偶然因素影响较大，且数据量较小，不足以对个体的特征进行稳健的估计。随着社交媒体、在线学习平台的普及，研究者能够收集到更加开放、更长时间跨度的文本数据，从而提取出感兴趣的、有意义的信息，揭示个体的心理特质(Dörre et al., 1999; Feldman & Sanger, 2006)。剑桥心理测量中心的研究团队基于

Facebook 用户发布的文本内容，借助 LIWC 等词典提取语言特征，预测用户的大五人格(Markovikj et al., 2013)；微软研究团队基于 Facebook 用户发布的文本内容和大五人格理论，构建基于社交媒体的人格预测系统(Bachrach et al., 2014)；Aung 和 Myint (2019)采用 LIWC、SPLICE 两个词典的特征以及用户的社交网络特征，预测 Facebook 用户的大五人格；Marouf 等人(2019)采用 LIWC 词典特征，应用朴素贝叶斯、决策树、随机森林、线性回归以及支持向量机多种分类模型，预测 Facebook 用户的神经质人格。我国研究者多依托微博平台对个体的特质、心理状态进行预测，譬如，利用中文版 LIWC 提取特征，检测用户发布的内容中所表达的心理状态(汪静莹 等, 2016)，识别对大五人格各维度具有预测作用的词汇(Gu et al., 2018; Qiu et al., 2017)，以及检测用户的抑郁、焦虑以及自杀倾向(Cheng et al., 2017)。

前述研究显示出文本挖掘技术应用于人格特质预测的有效性，为本研究利用文本挖掘技术预测羞怯奠定了方法学基础。此外，监测抑郁、焦虑等心理状态以及预测大五人格的研究也为本研究提供了理论基础：一方面，羞怯内在地表现为一种社交焦虑(Leary, 1986)，而文本挖掘是检测焦虑等心理状态的理想工具，相关研究结果能够为羞怯情绪的识别提供重要参考；另一方面，羞怯与内倾性和神经质等人格特质存在显著的相关(Hofstee et al., 1992; Jones et al., 2014; Kwiatkowska et al., 2019; La Sala et al., 2014; Sato et al., 2018)，内倾性在羞怯中表现为倾向于独处但能够应对必要的社交，而神经质在羞怯中表现为孤独感、低自尊以及对可能出现的尴尬情境的过度担忧(Eysenck, 1969)，因此，大五人格的文本挖掘结果有助于更好地理解和解释羞怯个体的文本特征，例如，高外倾性和高宜人性的个体倾向于使用更多的积极情绪词和社会历程词，人称代词“我们”与宜人性显著正相关(Gill et al., 2009; Mehl et al., 2006; Nowson, 2006; Oberlander & Gill, 2006; Qiu et al., 2012; Yarkoni, 2010)。

因此，本研究收集了小学生在线教学平台“教客学伴” (https://www.jiaokee.com/)上的作文、日记及长评论作为文本数据集，据此对小学生羞怯群体进行自动识别。该文本数据集具有较长的时间跨度，其内容、主题均有较高的自由度，能够充分反映小学生的日常学习生活。此外，已有研究发现存在不同类型的羞怯个体，比如，一些羞怯个体报告的消极想法较少但回避反应明显，一些羞怯个体报告的焦虑水平很高但几乎没有表现出行为困难，还有一些羞怯个体体验到强烈的负面情绪但生理反应不明显(Henderson et al., 2014)。因此，本研究将对羞怯的三个维度分别进行特征提取和模型构建，比较三个维度的语言风格和词汇使用特点，同时比较模型在三个维度上的预测精度。

2　研究过程

本研究采用文本分析方法来训练小学生羞怯的分类模型。首先，由小学生作答羞怯量表，基于量表分数将被试分为“羞怯群体”和“普通群体”，这个过程通常被叫做“打标签”，即对被试群体进行分类；其次，收集被试的在线写作文本，由于文本是非结构化的，计算机无法直接对其进行处理与分析，因而需要进行文本的向量化表征。本研究旨在关注小学生羞怯的词汇特征，即羞怯学生在日常写作文本中的词汇使用特点，因此，采用基于心理词典提取文本特征的方法，将每名学生的全部文本表征为一系列词频特征。然后，采用卡方算法来筛选重要特征。最后，采用机器学习算法，基于筛选后的特征构建小学生羞怯的预测模型。

2.1　数据收集

2.1.1　量表数据

在问卷星(https://www.wjx.cn/)上发布《国小儿童害羞量表》，邀请“教客学伴”平台上 2~5 年级的小学生在线作答量表，共回收问卷 2734 份，其父母在线签署了知情同意书。量表分为羞怯行为、羞怯认知、羞怯情绪三个维度，共 29 道题目，均采用 4 点计分，“1”表示“非常符合”，“2”表示“较符合”，“3”表示“较不符合”，“4”表示“非常不符合”。量表题目示例：羞怯行为“同学邀请我参加活动的时候，我经常找借口拒绝”；羞怯认知“如果我拒绝了别人的请求，我认为他们一定会对我产生不好的看法”。羞怯情绪“老师叫我回答问题的时候，不管我会不会，我都会心跳加快或手心冒汗”。

采用 SPSS 23.0 软件对量表数据进行清理。根据学生的作答用时以及总题量分析，剔除了用时过短(平均每题用时小于 2 秒)、全卷作答相同选项数量大于 70%，以及在相同数量的正反向题目上得分标准差小于 0.4 (说明被试没有考察题目的语义，无论正向或反向题目均给予了一致的回答)的无效被试，得到有效问卷 2476 份，男生 1284 名，女生 1192 名；各年级人数分布为：二年级 937 人，三年

级 1012 人，四年级 449 人，五年级 78 人。

基于上述 2476 名学生数据，采用 Mplus 7.0 软件对量表进行验证性因素分析及修订。基于修订后的量表，分别计算被试在 3 个维度上的原始得分，个体的量表得分越高表示其羞怯水平越高，随后将原始得分转换成标准分数。依据标准分数对个体进行分类是一种常用方法，比如，儿童行为评估系统(Behavior Assessment System for Children Third Version)将标准分大于 2 (即原始得分高于均值 2 个标准差以上)的个体判为“存在严重问题”，标准分大于 1 小于 2 (得分高于均值 1 个标准差到 2 个标准差之间)的个体判为“有风险” (Sandoval & Echandia, 1994)。本研究将标准分数大于 1 的个体划入“羞怯组” (标签为 1)，表示个体在该维度上表现出羞怯，将标准分数小于等于 1 的个体划入“普通组(标签为 0)，表示个体的羞怯特征不明显。各维度羞怯组与普通组人数比例为，羞怯行为 281 : 1026，羞怯认知 176 : 1131，羞怯情绪 217 : 1090。

2.1.2 文本数据

依据填写量表的学生 ID，收集学生在“教客学伴”上 2013 年 6 月至 2018 年 1 月的所有文本数据。该平台为语文教改背景下的实验教学平台，语文老师每天上课都会使用该教学平台，学生在语文课上经常被要求在平台上写作文，写作频率很高。除老师要求的写作任务外，学生还会写日记、对同学的写作进行评论，这部分内容比较少。本研究将上述文本数据合并在一起，共同作为文本数据来挖掘语言特征。由于小学生的写作形式仅限于记叙文，写作内容也多围绕小学生的真实经历，例如，记叙国庆假期的趣事、你最好的朋友等主题，因而合并后的文本数据记录了学生日常生活学习的方方面面，话题非常丰富，能够充分地反映出学生的语言风格和表达方式。

对上述文本数据进行清理，删去引用文本(如摘抄等)、无意义文本(如乱码等)、短文本(“已阅”等)以及重复文本，仅保留学生原创的作文、日志以及长评论。随后，将同一 ID 的文本进行汇总，剔除文本总量过低的个体。最后，将量表数据与文本数据的进行匹配，得到有效被试 1306 人。卡方检验结果显示，是否能够匹配与性别之间的卡方值不显著($\chi^2 = 1.19$, $df=1$, $p = 0.552$)，即与性别独立；各年级学生的文本数据缺失情况(能够匹配的人数 : 不能够匹配的人数)：二年级 416 : 521，三年级 614 : 398，四年级 229 : 220，五年级 47 : 31。这一分布主要是由于低年级在线写作任务较少。

学生文本示例：“有一次，我在书法课里，我正在和一个朋友上课，突然，老师说了一句话，说：明天就是书法竞赛，谁要参加？我们纷纷议论，大家都异口同声说：不想去。老师想：机会难得，只有两个名额，然后老师说了出来，我们又推人出来，而我静静地坐在教室里的一角，认认真真地写字，后来，老师大发雷霆，火冒三丈地说：大家都回座位。我们就坐好了。老师说这样就让，老师还没说出来，同学们都响起了掌声，说着我。我有点害羞了，心砰砰地跳个不停，那掌声真让我入迷，我都说不出来了。老师说：那就××去吧，还有他后面的那个谁一起去参加比赛。我很紧张当时，然后就让我回家好好练，最后就推选了我，还有我的朋友，那一次真让我难忘，我不会忘记的。”

2.2 心理词典修订与特征提取

在文本分析领域，词典是指定义了特定类别的单词的集合，包含词语归属的类别名称以及词列表。目前在心理学领域应用较广的是上世纪 90 年代开发的“语言探索与字词计数”词典(Linguistic Inquiry and Word Count, LIWC) (Pennebaker et al., 2015)，它包含 80 个词类，约 4500 个字词，已被用来研究人格特质、注意指向、思维方式、亲密关系、社会关系、情绪与心理健康等众多问题(Tausczik & Pennebaker, 2010)。将文本中的词语与词典进行一一比对，输出各类词语的词频结果，这是一个将文本向量化的过程，叫做特征提取(Tausczik & Pennebaker, 2010)。

研究者认为，基于词典的方法适用于聚焦特定的研究问题或主题的任务(Guo et al., 2016)。因此，词典的适用性对研究结果具有至关重要的影响，如果词典本身涵盖的词类不适于进行目标构念(羞怯)的分析，特征提取的有效性将会被削弱。本研究选择中科院心理所计算网络心理实验室研发的“文心(TextMind)”中文心理分析系统，该系统的核心词典(下称文心词典)参照 LIWC 开发，词库分类体系也与 LIWC 兼容一致(朱廷劭, 2016)。文心词典包含 102 个词类，超过一万个词，在心理学研究中应用广泛，其有效性得到了充分的验证(Lin et al., 2018; Shen et al., 2018; Wan et al., 2019; 赵楠 等, 2020)。本研究基于小学生羞怯的语言特征，对文心词典进行了扩充和改编：情绪是羞怯特质的重要表现指标，文心词典中对情感历程词的分类比较粗糙，本研究参照大连理工大学开发的中文情感词汇本体库中

的情绪词类别, 在文心词典的正向情绪词与负向情绪词下补充 7 个二级子类, 包括“乐”、“好”、“怒”、“哀”、“惧”、“恶”、“惊”; 将原词典的“生气词”、“悲伤词”与中文情感词汇本体库中的“愤怒”、“悲伤”合并为“怒”、“哀”两个类别, 将原词典的“焦虑词”归入“惧”这一类别下。此外, “动物词”在小学生文本中的出现频率较高, 将其补充到词典中; 最后, 删去中文分析时不常用的“冠词”类, 分别合并了过去、现在、将来时态下的多个子类。修订后的文心词典共包含 118 个类别。

此外, 文心词典中的词汇贴近成人语言, 而小学生的语言与成人存在较大差异。为保证词典适用于提取小学生文本特征, 本研究对收集到的“教客学伴”上的小学生文本进行汇总, 统计所有词汇的词频并由高到低排序, 将小学生使用频率较高但未包含于原文心词典的词汇纳入词典, 对各词类下的词汇进行扩充。

对每个小学生的全部文本进行分词后, 基于修订后的词典, 统计每个学生的文本在 118 个类别上的词汇频率, 从而生成 118 个文本特征变量。

2.3 特征筛选

采用文本分析来预测人格的相关研究大多预先进行特征筛选(如, Tadesse et al., 2018; Tandera et al., 2017; Yuan et al., 2018), 挑选重要的特征构建预测模型, 有利于提高模型的计算效率和预测准确性。本研究不仅构建预测模型, 而且着力于识别羞怯学生的日常用语与普通学生的差异从而刻画小学生羞怯的语言风格, 因此, 筛选出能够有效区分羞怯群体和普通群体的词汇进行分析是一个关键步骤。

首先, 对基于词典提取的 118 个特征进行初步筛选, 剔除与研究目的无关的特征。C1~C98 为 98 个词类, 这些词类之间存在层级关系, 例如, 感知历程词类别下包含视觉词、听觉词、感觉词; 情感历程词类别下包含正向情绪词以及负向情绪词。本研究认为不同层级的词类均有研究意义, 因此 C1~C98 均作为特征纳入模型。C99~C109 为 10 种标点符号, 本研究重点关注小学生羞怯的词汇特征, 因此 C99~C109 不纳入模型。C110~C118 为系统自动生成的统计特征, C110 为“总词数”, 由于学生开始使用平台进行在线写作的时间不同, 总词数相差较大, 因此不纳入模型, 仅利用该特征对 C1~C98 的词频进行标准化; C111 为“每句平均词数”, 我们认为这一特征能够反映出羞怯个体说话长短的语言特征, 因此将该特征纳入模型; C112 为“词典覆盖率”, 即该个体文本用词对整个词典的覆盖程度, 这一特征无法反映出羞怯群体与普通群体具体在哪一词类上存在差异, 因此不纳入模型; C113“数字比率”、C114“词长等于 4 的比率”、C115“词长大于 4 的比率”、C116“英文比率”、C118“URL 数量”, 均与本研究内容无关, 且小学生文本中基本不包含数字和英文文本, 因此这部分特征也不纳入模型; C117“情感词比率”与 C37“情感历程词”对本研究的意义相同, 因此前者不纳入模型。

然后, 采用卡方算法(Oakes et al., 2001)针对 C1~C98 以及 C111 共计 99 个词频特征进行进一步的筛选, 识别特征集合中能够最大程度地区分羞怯群体与普通群体的特征。卡方算法是特征筛选的常用算法(Forman, 2003), 例如 Paudel 等人(2018)通过卡方算法提取特征以实现对推特用户发布内容的情感分析。卡方算法在锁定文本中具有代表性的关键词等方面具有很高的有效性, 通过确定最小的特征集合, 使模型仅利用部分重要特征达到理想的预测水平。例如, Agarwal 等人(2011)对推特文本进行情感分类时, 使用语义以及句子成分等大量文本特征达到 75.39%的分类准确性, Chamansingh 和 Hosein (2016)在此基础上仅增加了卡方算法提取特征这一步骤, 将模型的分类准确性提高至 78.07%, 并且大大降低了模型的计算时间和内存需求。

卡方算法检验词频特征在两个群体的文本中出现的频率差异, 并且考虑两类文本长度的影响。卡方值越大, 表示该类词汇在两类人群中的使用频率差异越大。具体计算公式为:

$$\chi^2 = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11}+O_{12})(O_{11}+O_{21})(O_{12}+O_{22})(O_{21}+O_{22})}$$

其中, O_{11} 表示第 i 类词(W_i)在羞怯群体文本中出现的频次, O_{12} 表示 W_i 在普通群体文本中出现的频次; O_{21} 表示除 W_i 外的其它词($\neg W_i$)在羞怯群体文本中出现的频次, O_{22} 表示 $\neg W_i$ 在普通群体文本中出现的频次, N 表示训练文本中所有词语的频次总和, $N = O_{11} + O_{12} + O_{21} + O_{22}$。

本研究基于量表中羞怯行为、羞怯认知以及羞怯情绪三个维度的得分, 分别将小学生分为存在行为羞怯、认知羞怯和情绪羞怯的个体以及普通个体, 以 $\chi^2 > 100$ 作为筛选标准, 得到羞怯行为、羞怯认知以及羞怯情绪上的三组典型特征, 这部分特征除以总词数并标准化后, 纳入机器学习预测模型。

2.4 建立模型

基于筛选后的特征, 采用机器学习算法构建文

本特征对羞怯类别的预测模型。机器学习算法可以构建高维预测变量及非线性的预测模型。首先，将数据集分为训练集和测试集，基于训练集来训练机器学习模型，达到可接受的预测效果后，在测试集中评估模型预测的准确性。训练集和测试集的划分比例见表 1。

表 1 训练集及测试集划分

样本数	羞怯行为		羞怯认知		羞怯情绪	
	是	否	是	否	是	否
训练样本数	138	907	228	817	176	869
测试样本数	38	223	53	208	41	220

模型评估指标为：准确率、召回率及 F1 值。结合表 2 对各个指标进行解释：准确率表示被模型分到某类别的个体中，实际属于该类的比例。例如，模型在羞怯群体上的准确率表示被分类为羞怯的学生中，实际也为羞怯(标签为 1)的比例，公式为 TP/(TP+FP)；模型在普通群体上的准确率表示被分类为普通的学生中，实际也为普通(标签为 0)的比例，公式为 TN/(FN+TN)；两个准确率的均值代表总准确率。召回率表示实际属于某类别的个体中，被模型正确分到该类的比例。例如，模型在羞怯群体上的召回率表示实际为羞怯的学生中，被正确分类为羞怯的比例，这一指标也被称为“敏感度”，体现出该工具将“阳性”个体检测出来的有效性，公式为 TP/(TP+FN)；模型在普通群体上的召回率表示实际为普通的学生中，被正确分类为普通的比例，这一指标也被称为“特异度”，体现出工具将“非阳性”个体拒绝掉的有效性，公式为 TN/(FP+TN)；两个召回率的均值代表总召回率。F1 值为准确率与召回率的调和平均数，是模型的综合指标，羞怯群体的 F1 值为羞怯群体的准确率和召回率的调和平均数，普通群体的 F1 值为普通群体的准确率和召回率的调和平均数，总 F1 值为总准确率和总召回率的调和平均数。

表 2 模型分类结果及数据的实际分布

	实际属于羞怯群体	实际属于普通群体
分类后属于羞怯群体	TP	FP
分类属于普通群体	FN	TN

本研究使用 Python 3.6.2 进行模型构建，尝试采用决策树(Decision Tree, DT)、随机森林(Random Forest, RF)、支持向量机(Support Vector Machine, SVM)、逻辑斯蒂克回归(Logistics Regression, LR)、K 近邻(K-Nearest Neighbor, KNN)以及多层感知机(Multi-Layer Perceptron, MLP)六种分类模型，上述模型在人格相关研究中应用广泛并表现出较高的有效性(Aung & Myint, 2019; Farnadi et al., 2013; Majumder et al., 2017; Marouf et al., 2019)。

本研究中羞怯群体的比例远远小于普通群体，这种“样本类别分布不均衡”的情况普遍存在于心理学相关研究中，对模型预测的有效性具有负面影响(He & Garcia, 2009)。原始的分类器，如逻辑斯蒂克回归、决策树和支持向量机，均适用于均衡的数据集，在样本类别分布不均衡问题上表现不佳(López et al., 2013)。研究发现，在应用领域，构建损失敏感模型的结果往往优于其它方法(Liu & Zhou, 2006; McCarthy et al., 2005)，该方法通过设定错误分类的损失矩阵使模型达到最优预测效果(Ting, 2002)。本研究的主要目的是筛选出羞怯学生，因此为羞怯样本的错误分类设定更高的损失，并对各个模型进行独立调参，使模型达到最优的分类结果。

最后，综合比较各模型的预测结果，选择最优模型进行交叉验证。交叉验证的意义在于规避由于单次抽取的训练集和测试集导致的偏差，该方法能够对模型的泛化能力进行可靠的估计，评估模型的稳定性(Hawkins et al., 2003)。交叉验证结果较好，表明预测模型对个体的分类是稳定的，这与测量信度的概念较为一致。本研究采用 Geisser (1975)提出的 V 折交叉验证方法，V 设置为 5，即将数据平均分为 5 份，每次以其中 1 份作为测试集，以其余的 4 份作为训练集，进行多次验证。已有研究结果表明当 V 取值为 5~10 时，能够在达到验证效果稳定性的同时，保证模型计算的高效性(Friedman et al., 2001)。

3 研究结果

3.1 教师评分与学生自评的相关

预研究选择华西小学三年级 2 班的 49 名学生作为样本，收集学生在《国小儿童害羞量表》上的自评结果以及班主任对学生羞怯水平的评定结果。相关分析显示(表 3)，教师评分与学生自评之间相关较低。

3.2 小学生羞怯量表修订

首先，剔除题总相关小于 0.2 的 3 道题目(5.我会主动和刚认识的人说话；7.在班上，我经常主动举手回答问题；27.老师或长辈跟我说话的时候，我

表 3 教师评分与学生自评的相关

变量	教师评分–羞怯行为	教师评分–羞怯认知	教师评分–羞怯情绪	教师评分–总分
学生自评–羞怯行为	0.205	0.514	0.419	0.415
学生自评–羞怯认知	−0.170	0.213	−0.092	−0.008
学生自评–羞怯情绪	0.067	0.332	0.084	0.181
学生自评–总分	0.058	0.421	0.161	0.239

不会紧张)。随后，对羞怯量表进行验证性因素分析。其中部分题目存在较为明显的语义重叠，对这部分题目进行精简(例如：20.我和不熟悉的人在一起的时候，我会发抖或心跳加速；21.如果我和很多不认识的人在一起，我会感到身体不舒服；22.如果陌生人跟我说话，我会心跳加快；26.如果我和不熟悉的人在一起，我会感到紧张不安)。最终量表包含19 道题目，模型拟合良好(χ^2 = 898.599, *df* = 149, RMSEA = 0.04, CFI = 0.94, TLI = 0.94, SRMR = 0.04)。羞怯行为维度、羞怯认知维度、羞怯情绪维度以及量表整体的内部一致性(Cronbach's α 系数)分别为 0.79、0.66、0.78 以及 0.86。

3.3 描述统计

根据问卷的作答结果标定样本在羞怯行为、羞怯认知和羞怯情绪三个维度上的标签，个体维度总分 Z 分数大于 1 视为该维度的羞怯个体(标签为 1)，其余个体为普通个体(标签为 0)，由此划分出 6 组样本，各组样本量、词频等数据如表 4 所示。

表 4 描述统计结果汇总

维度	样本量	样本比例	最低词汇总量	最高词汇总量	词汇总量均值	词汇总量标准差
羞怯行为						
是(1)	281	21.52%	62	353086	7338.62	21711.80
否(0)	1025	78.48%	78	292321	10114.79	31150.93
羞怯认知						
是(1)	176	13.48%	137	152657	6056.41	12702.45
否(0)	1130	86.52%	62	353086	8227.89	25375.75
羞怯情绪						
是(1)	217	16.62%	95	199348	7957.13	21606.92
否(0)	1089	83.38%	62	353086	7931.17	24538.95

3.4 心理词典修订

汇总“教客学伴”上所有学生的文本数据，使用哈工大社会计算与信息检索研究中心研发的语言技术平台(LTP)对文本进行自动分词，统计每个词出现的频次，将所有词频由高到低进行排序，并且将词频逐个累加得到累计词频率。结果显示，前6000 个词的累积词频率超过 90%，此后每再增加1000 个词，累积词频率的增量均不超过 1% (累积词频率和累积词频率增量见图 1 图 2)。因此，保留前 6000 个词作为小学生文本的高频词。将 6000 个高频词与原词典中的词汇进行比较，其中 3119 个词能够完全匹配，其余 2881 个词根据词义分别归入 118 个词类，该归类任务由两名心理学研究生协商共同完成。扩充后的词典详见 https://xttian.oss-cn-beijing.aliyuncs.com/shyness dictionary。

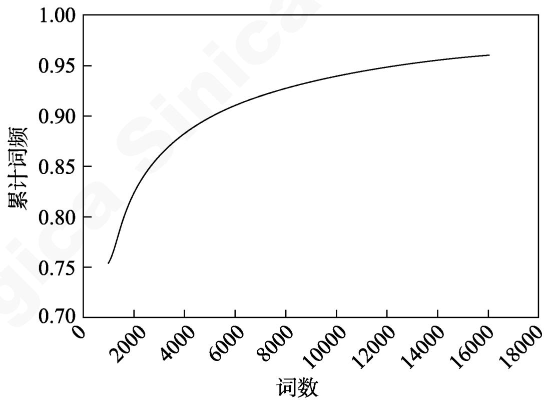


图 1 小学生日志文本累计词频

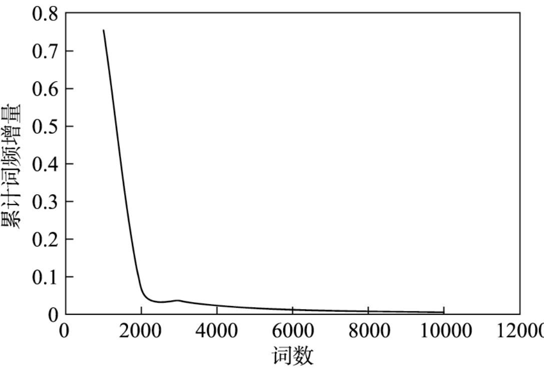


图 2 小学生日志文本累计词频增量

3.5 特征提取

基于修订后的心理词典，计算羞怯群体文本以及普通群体文本在各个词类上的词频。采用卡方算法，以卡方值大于 100 为标准，挑选出在两类群体文本中存在明显的频率差异的词类。羞怯行为共计提取 23 个词频特征，羞怯认知共计提取 16 个词频特征，羞怯情绪共计提取 20 个词频特征。将特征按“多维度共有”或“单一维度特有”的原则整理，结果显示：1)每句平均词数对羞怯行为、羞怯认知和羞

怯情绪均有负向预测作用；2)社会历程词对羞怯认知、羞怯行为以及羞怯情绪均具有负向预测作用；3)羞怯个体文本中的各类人称代名词均较少，其中，第一人称复数代名词对羞怯行为具有正向预测作用，而对羞怯认知和羞怯情绪具有负向预测作用；4)口语词、功能词、认知历程词等词类对羞怯行为具有正向预测作用，而对羞怯认知和羞怯情绪具有负向的预测作用；5)时间词中的过去词对羞怯行为具有正向预测作用，而延续词则对羞怯情绪有明显的负向预测作用；6)感知历程词，尤其是视觉词，对羞怯行为具有明显的正向预测作用；7)生理历程词，主要是摄食词，对羞怯情绪具有明显的负向预测作用。

各维度特征及卡方值详见表 5。

3.6 模型预测

基于筛选后的特征，分别对羞怯行为、羞怯认知、羞怯情绪建立预测模型。本研究使用 6 种分类模型均在原始模型中加入损失敏感函数。比较不同模型的预测效果(见表 6)发现，对于普通群体的准确率，各模型间差异不大，均在 0.8 左右；对于羞怯群体的准确率，随机森林和逻辑斯蒂克回归的表现相对较好(0.15~0.50)；对于普通群体的召回率(即特异度)，逻辑斯蒂克回归和 K 近邻的表现相对较好(大于 0.9)；对于羞怯群体的召回率(即敏感度)，随机森林和支持向量机的表现较好(0.15~0.44，其中情绪维度较高，行为和认知维度较低)；对于总 F1 值，随机森林的结果表现最好(0.55~0.57)。

结果还表明，模型对普通群体预测的准确率和召回率比较高，对羞怯群体预测的准确率和召回率比较低，这与羞怯群体的人数较少，难以从大群体中识别出来有关，虽然在原始模型中增加了损失敏感函数来提高羞怯群体的预测准确率，但是提升效果有限。

整体来看，随机森林模型在本研究任务中的表现最优，进而对该模型进行交叉验证，结果见表 7。结果显示，总 F1 的稳定性非常高，取值在 0.52~0.57

表 5 各维度提取特征

羞怯行为				羞怯认知				羞怯情绪			
词类	含义	方向	$\chi^2(1)$	词类	含义	方向	$\chi^2(1)$	词类	含义	方向	$\chi^2(1)$
C111	每句平均词数	−	534.36	C111	每句平均词数	−	534.36	C111	每句平均词数	−	534.36
C33	社会历程词	−	159.70	C33	社会历程词	−	579.58	C33	社会历程词	−	293.28
C34	家人词	−	377.82	C34	家人词	−	274.53	C34	家人词	−	236.62
C28	口语词	+	146.82	C28	口语词	−	300.67	C28	口语词	−	131.76
C32	填充赘词	+	143.77	C32	填充赘词	−	331.44	C32	填充赘词	−	100.69
C67	知历程词	+	421.76	C67	认知历程词	−	301.68	C67	认知历程词	−	270.85
C5	第一人称复数代名词	+	152.95	C5	第一人称复数代名词	−	406.14	C3	特定人称代名词	−	132.93
C6	第二人称代名词	−	276.34	C2	代名词	−	256.97	C74	包含词	−	225.94
C31	停顿赘词	+	163.20	C3	特定人称代名词	−	334.49	C69	因果词	+	112.08
C72	确切词	+	204.60	C74	包含词	−	121.67	C1	功能词	−	270.38
C75	排除词	+	144.94	C75	排除词	−	243.69	C22	语助词	−	116.55
C69	因果词	+	228.00	C15	连接词	−	126.65	C11	动词	−	334.79
C1	功能词	+	359.20	C90	成就词	+	118.25	C91	休闲词	−	271.43
C13	副词	+	315.86	C85	相对词	+	104.50	C23	时态标定词	−	319.35
C15	连接词	+	574.02	C87	空间词	+	226.83	C27	延续	−	334.47
C91	休闲词	−	307.10	C94	宗教词	−	332.24	C85	相对词	−	424.11
C98	动物词	−	477.33					C88	时间词	−	174.19
C24	过去	+	113.93					C86	移动词	−	367.58
C89	工作词	+	270.43					C80	生理历程词	−	111.72
C76	感知历程词	−	340.46					C84	摄食词	−	126.11
C77	视觉词	−	309.74								
C78	听觉词	−	204.74								
C94	宗教词	+	139.86								

注：+表示该特征在羞怯群体文本中出现的频率更高，−表示在普通群体文本中出现的频率更高。

表 6 模型预测结果

维度	决策树			随机森林			逻辑斯蒂克回归			K 近邻			多层感知机			支持向量机		
	行为	认知	情绪	行为	认知	情绪	行为	认知	情绪	行为	认知	情绪	行为	认知	情绪	行为	认知	情绪
准确率–普通	0.86	0.78	0.85	0.88	0.84	0.87	0.86	0.80	0.84	0.85	0.79	0.84	0.83	0.81	0.85	0.84	0.82	0.87
准确率–羞怯	0.16	0.16	0.18	0.29	0.25	0.20	0.50	0.27	0.15	0.00	0.14	0.20	0.10	0.27	0.16	0.11	0.28	0.23
召回率–普通	0.86	0.75	0.82	0.86	0.93	0.68	0.99	0.92	0.90	0.99	0.94	0.98	0.63	0.89	0.67	0.75	0.78	0.74
召回率–羞怯	0.16	0.19	0.22	0.29	0.15	0.44	0.05	0.11	0.10	0.00	0.04	0.02	0.24	0.17	0.34	0.18	0.34	0.42
F1 值–普通	0.86	0.77	0.83	0.89	0.67	0.76	0.92	0.86	0.87	0.00	0.86	0.91	0.71	0.85	0.75	0.79	0.80	0.80
F1 值–羞怯	0.16	0.17	0.20	0.28	0.34	0.28	0.10	0.16	0.12	0.00	0.06	0.04	0.14	0.21	0.22	0.14	0.31	0.29
总 F1 值	0.51	0.47	0.52	0.57	0.56	0.55	0.59	0.53	0.50	0.46	0.48	0.51	0.45	0.53	0.51	0.47	0.56	0.56

表 7 随机森林模型交叉验证结果

维度	第一次			第二次			第三次			第四次			第五次		
	行为	认知	情绪	行为	认知	情绪	行为	认知	情绪	行为	认知	情绪	行为	认知	情绪
准确率–普通	0.88	0.81	0.87	0.85	0.75	0.86	0.89	0.78	0.86	0.88	0.83	0.86	0.89	0.79	0.85
准确率–羞怯	0.26	0.35	0.20	0.20	0.43	0.26	0.18	0.27	0.20	0.19	0.33	0.19	0.22	0.29	0.18
召回率–普通	0.86	0.93	0.68	0.87	0.96	0.76	0.7	0.87	0.77	0.73	0.85	0.69	0.74	0.82	0.62
召回率–羞怯	0.29	0.15	0.44	0.17	0.09	0.4	0.43	0.17	0.33	0.39	0.29	0.39	0.44	0.25	0.44
F1 值–普通	0.87	0.87	0.76	0.86	0.84	0.81	0.78	0.82	0.81	0.80	0.84	0.77	0.81	0.80	0.72
F1 值–羞怯	0.27	0.21	0.28	0.18	0.15	0.32	0.25	0.21	0.25	0.26	0.31	0.26	0.29	0.27	0.26
总 F1 值	0.57	0.56	0.55	0.52	0.56	0.57	0.55	0.52	0.54	0.55	0.57	0.53	0.57	0.54	0.52

之间。普通群体的准确率和召回率都比较稳定，波动最大不超过 0.17，大部分的波动都不超过 0.1。羞怯群体的稳定性相对差一些，波动最大为 0.27，大部分波动都不超过 0.15。

4 讨论

本研究旨在探索小学生羞怯的语言风格，关注羞怯群体与普通群体的词汇使用差异，因此预先采用卡方算法进行词频特征的筛选，再基于筛选后的特征来构建预测模型。与此相反，有一些研究以模型准确率为主要目标，不强调模型结果的解释性，采用神经网络等更加复杂的深度模型处理高维特征，而不进行特征筛选(如，Xue et al., 2018)，采用该方式建模虽然预测精度较高，但是并不能够对预测群体的语言风格进行直观的描述和解释。因而，采用卡方算法预先提取特征再构建预测模型更符合本研究的目标。下面分别从羞怯群体的词汇使用的共性和差异性以及羞怯预测模型的精度展开讨论。

4.1 羞怯各维度的文本表现既有共性也有特性

本研究基于真实的小学生文本对“文心词典”进行扩充，构建出适用于儿童的心理词典，实现更为精准的特征提取，并通过卡方检验对词频特征进行筛选。结果显示，羞怯群体和普通群体在用词上确实存在明显差异。部分特征为多个维度所共有，即羞怯行为、羞怯认知和羞怯情绪的普遍的语言特点；同时也存在某些特征为单一维度所有，能够充分体现出该维度的特点。

羞怯行为、羞怯认知以及羞怯情绪的文本特征的共性主要体现在“每句平均词数”以及“社会历程词”两方面：首先，高羞怯行为、高羞怯认知以及高羞怯情绪个体的文本中，每句平均词数均小于普通个体文本，这与羞怯个体说话少、回答较短、表达不流利等行为模式有直接关系。已有研究显示，高内倾性个体的每句平均词数比外倾性个体更少(Mehl et al., 2006)，而羞怯的青少年更有可能是内倾性人格(Lawrence & Bennett, 1992)，且越内倾的个体羞怯感越强(韩磊 等, 2011)。因此，羞怯个体的语言风格与高内倾性个体相近，在文本中体现为使用更短的句子；其次，高羞怯认知、高羞怯行为以及高羞怯情绪个体的文本中，社会历程词均较少。已有研究显示，高外倾性的个体文本中的社会性词汇更多(Gill et al., 2009; Mehl et al., 2006)，因此，较少的社会历程词进一步体现出羞怯个体的内倾性。社会历程词对羞怯的负向预测作用主要体现在家人词这一子类，高羞怯个体的文本中更少提及

家人词。已有研究显示，儿童期羞怯与亲子关系、父母的社交能力等均有关系(Arroyo et al., 2012; Huang, 1999)，羞怯儿童的父母往往对儿童表达的需求不敏感，且更倾向于采用强硬的策略(Engfer, 1993)，控制型或过度保护型的家庭涉及频繁的矫正和羞辱，更可能引起儿童羞怯(Bruch, 1989)。因此，尽管家人词是小学生文本中的高频词，但羞怯学生文本中这类词相对更少，反映出家庭环境对羞怯特质发展的影响。

已有研究者认为，羞怯的内在认知及情绪体验应当与其外在行为表现区分(Leary, 1986)。本研究对羞怯行为、羞怯认知和羞怯情绪三个维度分别提取特征，结果显示，羞怯行为在文本表现上确与羞怯认知及情绪存在较为明显的差异。首先，高羞怯认知和高羞怯情绪的个体更少提及“我们”等第一人称复数代名词，而高羞怯行为个体的文本中这类词汇则较多，可能反映出羞怯个体的趋避冲突(Coplan et al., 2004)，即高社交趋向动机和高社交回避动机的组合，尽管羞怯个体感到社交焦虑和压力，但同样期望进行社交活动以维持自尊水平(Dennissen et al., 2008)；其次，口语词(包含填充赘词等)、功能词(包括连接词等)、认知历程词(包含因果词、包含词、排除词等)对羞怯行为具有正向预测作用，而对羞怯认知和羞怯情绪具有负向预测作用。已有研究显示，较多的填充赘词反映出个体更多地考虑他人和环境，即对他人接受自己所表达内容的期待(Laserna et al., 2014)。上述特征提取结果反映出羞怯行为维度的独特性。另外，羞怯行为往往受到外部限制，具有较高的情境性(Henderson et al., 2014)，在写作文本中的特征体现可能不明显。

时间词包含过去、现在、未来以及延续四个子类，其中过去词对羞怯行为具有明显的正向预测作用，而延续词则对羞怯情绪有明显的负向预测作用。已有研究表明，高羞怯个体往往存在低自我认知、低自尊等问题(Caspi et al., 1988; Dennissen et al., 2008)。黄希庭和郑涌(2000)研究发现，高自我认同个体有更大的现在广度，更积极的未来取向，而低自我认同个体则有更强的过去取向，且与未来中断的现象更易发生。低自我认同者往往对现实感到无所适从、对未来感到彷徨，因此转向过去。因此，较多过去词、较少延续词的现象可能反映出羞怯个体的自我意识发展问题。

感知历程词对羞怯行为具有明显的正向预测作用，尤其体现在视觉词，这一结果与量表中的相关项目(如“我觉得别人在注意我”)相吻合。研究者普遍认为，羞怯包含个体对他人如何看待和评估自己的担忧，当个体关注他人对自己的印象时，往往会激起羞怯的相关反应(Leary & Schlenker, 1981; Leary, 1983; Schlenker & Leary, 1982)，较多的感知历程词体现出高羞怯行为个体在日常生活中频繁地感觉到被注视、被评价，充分反映出羞怯个体的过度自我中心以及过度关注社会评价的特点(Zimbardo, 1982)。

生理历程词对羞怯情绪具有明显的负向预测作用，尤其体现在摄食词。基于新浪微博的研究结果显示，摄取词与神经质呈显著正相关(Gu et al., 2018)，而羞怯与神经质也显著相关(Hofstee et al., 1992; Jones et al., 2014; Kwiatkowska et al., 2019; La Sala et al., 2014; Sato et al., 2018)。进食障碍个体往往存在社会依赖与自主的趋避冲突，渴望保持独立的同时也依赖人际关系以维持自尊(Narduzzi & Jackson, 2000)，并且存在对错误的过度关注和对自我表现的焦虑(Cassin & von Ranson, 2005)，这些特征均与羞怯个体相似。

4.2 基于文本特征构建分类模型能够预测小学生羞怯水平

基于卡方算法识别重要的词频特征，据此构建分类预测模型并计算模型的各个精度指标，在识别重要特征组合的同时，保证了模型的预测效果。结果显示，随机森林在三个维度的表现较为均衡。已有研究中，同样有研究者采用随机森林在人格预测、心理问题筛查等任务中达到最优效果(Chen et al., 2017; Kwiatkowska et al., 2019; Papamitsiou & Economides, 2017; Skowron et al, 2016)。

由于前人尚未有针对羞怯特质的文本挖掘研究，因此将已有的人格预测研究结果与本研究对比分析。已有基于 LIWC 提取特征预测 Facebook 用户的大五人格的研究中，模型最终准确率、召回率以及 F1 值往往在 45%~65%范围内(如, Farnadi et al., 2013; Marouf et al., 2019)，本研究结果接近已有研究结果，显示出基于在线写作文本预测小学生羞怯特质的可行性，尤其是模型的特异度很高(大于 0.9)，说明对普通群体的误判非常小。羞怯情绪维度的敏感度在 0.4 左右，其它两个维度在 0.2 左右，这个结果并不理想，这主要是由于本研究的主要目标为探索小学生羞怯在词汇层面的表现，模型仅纳入词频特征，未采用对语言风格不具有可解释性的文本向量化表征方式，因此预测力有限。已有研究

显示，采用多元类型的特征(如用户的在线社交网络)以及深度学习模型，能够明显提高预测效果(Aung & Myint, 2019; Farnadi et al., 2013; Tandera et al., 2017)，比如，Farnadi 等人(2013)在词汇特征的基础上加入社交网络特征，模型的预测准确率从54%提升至 71%。因此，未来将尝试在已有发现的基础上，结合学生的在线行为数据等特征，构建更加复杂且具有良好解释性的模型。此外，羞怯群体的准确率和召回率不高，而且交叉验证的结果也显示羞怯群体的稳定性相对差一些，这一情况主要是受到了样本类别分布不均衡的影响(Diamantidis et al., 2000; Oommen et al., 2011)，未来将持续收集文本数据，扩充文本数据库，提升模型的预测效果。

目前，学生在教学平台上学习并在线完成作业和任务逐渐成为常态，未来必将产生更大量级的数据，为研究者基于文本挖掘学生心理品质提供丰富的原始语料。心理学研究应当充分发挥这一数据资产的价值，来挖掘和揭示心理特点和规律。本研究旨在探索基于在线教学平台上的文本数据来揭示和预测心理特质的新方法，期望为未来更多相关研究的开展提供参考。

5 结论

本研究尝试利用文本挖掘技术对小学生在线写作文本进行分析，挖掘羞怯特质的词汇特征，构建小学生羞怯特质的语言风格模型并实现自动预测。本研究的主要结果如下：

(1)基于真实的小学生文本对“文心词典”进行扩充，构建出适用于儿童的心理词典，实现更为精准的特征提取，并利用卡方算法筛选文本特征，构建出羞怯特质的多维度语言风格模型，识别出羞怯在行为、认知及情绪上的文本表现的共性和差异；(2)基于筛选后的词典特征建立多种预测模型，显示出机器学习模型在检测小学生羞怯特质上的有效性。

本研究揭示了小学生羞怯行为、羞怯认知和羞怯情绪的语言风格和词汇使用特点，为研究者更加深入地了解羞怯个体提供建议和启发。同时，本研究采用词汇特征构建多种机器学习模型，为未来采用更丰富的特征以及深度学习模型预测羞怯提供了基础。

参 考 文 献

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. J. (2011 June). *Sentiment analysis of twitter data*. In Proceedings of the Workshop on Languages in Social Media (pp. 30–38).

Allport, G. W. (1937). Personality: A psychological interpretation. *American Journal of Sociology*, *45*(1), 48–50.

Argamon, S., Dhawle, S., Koppel, M., & Pennebaker, J. W. (2005 June). *Lexical predictors of personality type*. In Proceedings of the 2005 Joint Annual Meeting of the Interface and the Classification Society of North America (pp. 1–16).

Arroyo, A., Nevárez, N., Segrin, C., & Harwood, J. (2012). The association between parent and adult child shyness, social skills, and perceived family communication. *Journal of Family Communication*, *12*(4), 249–264.

Aung, Z. M. M., & Myint, P. H. (2019 July). *Personality prediction based on content of Facebook users: A literature review*. In 2019 20th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (pp. 34–38).

Azucar, D., Marengo, D., & Settanni, M. (2018). Predicting the big 5 personality traits from digital footprints on social media: a meta-analysis. *Personality and Individual Differences*, *124*, 150–159.

Baardstu, S., Coplan, R. J., Karevold, E. B., Laceulle, O. M., & von Soest, T. (2019). Longitudinal pathways from shyness in early childhood to personality in adolescence: do peers matter? *Journal of Research on Adolescence*, *30*(2), 362–379.

Bachrach, Y., Graepel, T., Kohli, P., Kosinski, M., & Stillwell, D. (2014 May). *Your digital image: factors behind demographic and psychometric predictions from social network profiles*. In Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems (pp. 1649–1650).

Bai, S. T., Hao, B. B., Li, A., Nie, D., & Zhu, T. S. (2014). Depression and anxiety prediction on microblogs. *Journal of University of Chinese Academy of Sciences*, *31*(6), 814–820.

[白朔天, 郝碧波, 李昂, 聂栋, 朱廷劭. (2014). 微博用户的抑郁和焦虑预测. *中国科学院大学学报*, *31*(6), 814–820.]

Brodt, S. E., & Zimbardo, P. G. (1981). Modifying shyness-related social behavior through symptom misattribution. *Journal of Personality and Social Psychology*, *41*(3), 437–449.

Bruch, M. A. (1989). Familial and developmental antecedents of social phobia: issues and findings. *Clinical Psychology Review*, *9*(1), 37–47.

Caspi, A., Elder, G. H., & Bem, D. J. (1988). Moving away from the world: Life-course patterns of shy children. *Developmental Psychology*, *24*(6), 824–831.

Cassin, S. E., & von Ranson, K. M. (2005). Personality and eating disorders: A decade in review. *Clinical Psychology Review*, *25*(7), 895–916.

Cattell, R. B. (1943). The description of personality. *Psychological Review*, *50*, 559–594.

Chamansingh, N., & Hosein, P. (2016 September). *Efficient sentiment classification of Twitter feeds*. In 2016 IEEE International Conference on Knowledge Engineering and Applications (pp. 78–82).

Cheek, J. M., & Buss, A. H. (1981). Shyness and sociability. *Journal of Personality and Social Psychology*, *41*(2), 330–339.

Chen, L. S., Gong, T., Kosinski, M., Stillwell, D., & Davidson, R. L. (2017). Building a profile of subjective well-being for social media users. *Plos One*, *12*(11), e0187278.

Chen, X. Y. (2010). Shyness-inhibition in childhood and

adolescence: A cross-cultural perspective. In K. H. Rubin & R. J. Coplan (Eds.), *The development of shyness and social withdrawal* (pp. 213–235). New York: Guilford Press.

Cheng, Q., Li, T. M., Kwok, C. L., Zhu, T., & Yip, P. S. (2017). Assessing suicide risk and emotional distress in Chinese social media: A text mining and machine learning study. *Journal of Medical Internet Research*, *19*(7), e243.

Coplan, R. J., Prakash, K., O'Neil, K., & Armer, M. (2004). Do you "want" to play? distinguishing between conflicted shyness and social disinterest in early childhood. *Developmental Psychology*, *40*(2), 244–258.

Dennissen, J. J. A., Asendorpf, J. B., & van Aken, M. A. G. (2008). Childhood personality predicts long - term trajectories of shyness and aggressiveness in the context of demographic transitions in emerging adulthood. *Journal of Personality*, *76*(1), 67–100.

Diamantidis, N. A., Karlis, D., & Giakoumakis, E. A. (2000). Unsupervised stratification of cross-validation for accuracy estimation. *Artificial Intelligence*, *116*(1-2), 1–16.

Dörre, J., Gerstl, P., & Seiffert, R. (1999 August). *Text mining: finding nuggets in mountains of textual data*. In Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 398–401).

Eisenberg, N., Shepard, S. A., Fabes, R. A., Murphy, B. C., & Guthrie, I. K. (1998). Shyness and children's emotionality, regulation, and coping: Contemporaneous, longitudinal, and across-context relations. *Child Development*, *69*(3), 767–790.

Engfer, A. (1993). Antecedents and consequences of shyness in boys and girls: A 6-year longitudinal study. In K. H. Rubin, & J. Asendorpf (Eds.), *Social withdrawal, inhibition, and shyness in childhood* (pp. 49–79). London: Psychology Press.

Evans, M. (2010). Language, performance, academic performance, and signs of shyness. In K. H. Rubin & R. J. Coplan (Eds.), *The development of shyness and social withdrawal* (pp. 172–212). New York: Guilford Press.

Eysenck, H. J., & Cookson, D. (1969). Personality in primary school children: Ability and achievement. *British Journal of Educational Psychology*, *39*(2), 109–122.

Farnadi, G., Zoghbi, S., Moens, M. F., & de Cock, M. (2013 July). *Recognising personality traits using Facebook status updates*. In Seventh International AAAI Conference on Weblogs and Social Media (pp.14–18).

Fast, L. A., & Funder, D. C. (2008). Personality as manifest in word use: correlations with self-report, acquaintance report, and behavior. *Journal of Personality and Social Psychology*, *94*(2), 334–346.

Feldman, R., & Sanger, J. (2006). *The text mining handbook: Advanced approaches in analyzing unstructured data*. Cambridge University Press.

Findlay, L. C., Coplan, R. J., & Bowker, A. (2009). Keeping it all inside: shyness, internalizing coping strategies and socio-emotional adjustment in middle childhood. *International Journal of Behavioral Development*, *33*(1), 47–54.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, *3*, 1289–1305.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning*. New York: Springer series in statistics.

Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association*, *70*(350), 320–328.

Gill, A. J., Nowson, S., & Oberlander, J. (2009 March). *What are they blogging about? Personality, topic and motivation in blogs*. In Third International AAAI Conference on Weblogs and Social Media.

Goldberg, L. R. (1982). From Ace to Zombie: some explorations in the language of personality. *Advances in Personality Assessment*, *1*, 203–234.

Gu, H. Q., Wang, J., Wang, Z. W., Zhuang, B. J., & Su, F. (2018 July). *Modeling of user portrait through social media*. In 2018 IEEE International Conference on Multimedia and Expo (pp. 1–6).

Guo, L., Vargo, C. J., Pan, Z. X., Ding, W. C., & Ishwar, P. (2016). Big social data analytics in journalism and mass communication: comparing dictionary-based text analysis and unsupervised topic modeling. *Journalism and Mass Communication Quarterly*, *93*(2), 332–359.

Han, L., Gao, F. Q., Guo, Y. Y., & Wang, P. (2011). The relationships between personality and shyness: Mediation and moderation effects. *Journal of Psychological Science*, *34*(4), 889–893.

[韩磊, 高峰强, 郭永玉, 王鹏. (2011). 人格与羞怯的关系: 中介效应与调节效应. *心理科学*, *34*(4), 889–893.]

Hawkins, D. M., Basak, S. C., & Mills, D. (2003). Assessing model fit by cross-validation. *Journal of Chemical Information and Modeling*, *43*(2), 579–586.

He, H. B., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284.

Henderson, L., Gilbert, P., & Zimbardo, P. (2014). Shyness, social anxiety, and social phobia. In S. G. Hofmann & P. M. DiBartolo (Eds.), *Social anxiety* (pp. 95–115). London: Academic Press.

Hirsh, J. B., & Peterson, J. B. (2009). Personality and language use in self-narratives. *Journal of Research in Personality*, *43*(3), 524–527.

Ho, D. Y. F. (1986). Chinese patterns of socialization: a critical review. In M. H. Bond (Ed.), *The psychology of the Chinese people* (pp. 1–37). New York: Oxford University Press.

Hofstee, W. K., de Raad, B., & Goldberg, L. R. (1992). Integration of the big five and circumplex approaches to trait structure. *Journal of Personality and Social Psychology*, *63*(1), 146–163.

Huang, L. N. (1999). Family communication patterns and personality characteristics. *Communication Quarterly*, *47*(2), 230–243.

Huang, X. T., & Zheng, Y. (2000). Self-integration on time perspective: I. projective test for the psychological structure. *Acta Psychologica Sinica*, *32*(1), 30–35.

[黄希庭, 郑涌. (2000). 时间透视的自我整合: I.心理结构方式的投射测验. *心理学报*, *32*(1), 30–35.]

Jones, K. M., Schulkin, J., & Schmidt, L. A. (2014). Shyness: subtypes, psychosocial correlates, and treatment interventions. *Psychology*, *5*(3), 244–254.

Kagan, J. (1997). Temperament and the reactions to unfamiliarity. *Child Development*, *68*(1), 139–143.

Karevold, E., Ystrom, E., Coplan, R. J., Sanson, A. V., & Mathiesen, K. S. (2012). A prospective longitudinal study of shyness from infancy to adolescence: Stability, age-related changes, and prediction of socio-emotional functioning. *Journal of Abnormal Child Psychology*, *40*(7), 1167–1177.

Kwiatkowska, M. M., Jułkowski, T., Rogoza, R., Żemojtel-Piotrowska, M., & Fatfouta, R. (2019). Narcissism and trust: differential impact of agentic, antagonistic, and communal narcissism. *Personality and Individual Differences*, *137*,

139–143.

La Sala, L., Skues, J., & Grant, S. (2014). Personality traits and Facebook use: the combined/interactive effect of extraversion, neuroticism and conscientiousness. *Social Networking*, *3*(5), 211–219.

Laserna, C. M., Seih, Y. T., & Pennebaker, J. W. (2014). Um… who like says you know: filler word use as a function of age, gender, and personality. *Journal of Language and Social Psychology*, *33*(3), 328–338.

Lawrence, B., & Bennett, S. (1992). Shyness and education: the relationship between shyness, social class and personality variables in adolescents. *British Journal of Educational Psychology*, *62*(2), 257–263.

Leary, M. R. (1983). *Understanding social anxiety: Social, personality, and clinical perspectives*. Beverly Hills: Sage Publication.

Leary, M. R. (1986). Affective and behavioral components of shyness. In W. H. Jones, J. M. Cheek, & S. R. Briggs (Eds.), *Shyness* (pp. 27–38). Springer Science & Business Media.

Leary, M. R., & Schlenker, B. R. (1981). The social psychology of shyness: A self-presentation model. In J. T. Tedeschi (Ed.), *Impression management theory and social psychological research* (pp. 335–358). London: Academic Press.

Lewinsky, H. (1941). The nature of shyness. *British Journal of Psychology*, *32*(2), 105–113.

Li, Y. H., Hu, W. D., Xu, Z. P., & Han, W. Q. (2005). Study and progress of socially desirable responding problem in personality measurement. *Chinese Journal of Clinical Rehabilitation*, *9*(8), 119–121.

[李亚红, 胡文东, 徐志鹏, 韩文强. (2005).人格测量中社会期许式作答问题的研究与进展. *中国组织工程研究*, *9*(8), 119–121.]

Lin, B. Y., Xu, F. F., Zhu, K., & Hwang, S. W. (2018, July). *Mining cross-cultural differences and similarities in social media*. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (pp. 709–719).

Liu, X. Y., & Zhou, Z. H. (2006, December). *The influence of class imbalance on cost-sensitive learning: An empirical study*. In Sixth International Conference on Data Mining (pp. 970–974).

López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences*, *250*, 113–141.

Luo, F., & Zhang, H. C. (2007). Methods of coping with faking of personality tests. *Psychological Exploration*, *27*(4), 78–82.

[骆方, 张厚粲. (2007). 人格测验中作假的控制方法. *心理学探新*, *27*(4), 78–82.]

Mairesse, F., Walker, M. A., Mehl, M. R., & Moore, R. K. (2007). Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research*, *30*(1), 457–500.

Majumder, N., Poria, S., Gelbukh, A., & Cambria, E. (2017). Deep learning-based document modeling for personality detection from text. *IEEE Intelligent Systems*, *32*(2), 74–79.

Markovikj, D., Gievska, S., Kosinski, M., & Stillwell, D. (2013, June). *Mining Facebook data for predictive personality modeling*. In Seventh International AAAI Conference on Weblogs and Social Media (pp. 23–26).

Marouf, A. A., Hasan, M. K., & Mahmud, H. (2019 February). *Identifying neuroticism from user generated content of social media based on psycholinguistic cues*. In 2019 International Conference on Electrical, Computer and Communication Engineering (pp. 1–5).

Martin, B. (1961). The assessment of anxiety by physiological behavioral measures. *Psychological Bulletin*, *58*, 234–255.

Mccarthy, K., Zabar, B., & Weiss, G. (2005, August). *Does cost-sensitive learning beat sampling for classifying rare classes?* In International Workshop on Utility-based Data Mining (pp. 69–77).

Mehl, M. R., Gosling, S. D., & Pennebaker, J. W. (2006). Personality in its natural habitat: manifestations and implicit folk theories of personality in daily life. *Journal of Personality and Social Psychology*, *90*(5), 862–877.

Narduzzi, K. J., & Jackson, T. (2000). Personality differences between eating‐disordered women and a nonclinical comparison sample: A discriminant classification analysis. *Journal of Clinical Psychology*, *56*(6), 699–710.

Norman, W. T. (1963). Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *Journal of Abnormal and Social Psychology*, *66*(6), 574–583.

Nowson, S. (2006). *The language of weblogs: a study of genre and individual differences* (Unpublished doctoral dissertation). University of Edinburgh.

Oakes, M., Gaaizauskas, R., Fowkes, H., Jonsson, A., Wan, V., & Beaulieu, M. (2001, September). *A method based on the chi-square test for document classification*. In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 440–441).

Oberlander, J., & Gill, A. J. (2006). Language with character: A stratified corpus comparison of individual differences in e-mail communication. *Discourse Processes*, *42*(3), 239–270.

Oommen, T., Baise, L. G., & Vogel, R. M. (2011). Sampling bias and class imbalance in Maximum-likelihood Logistic Regression. *Mathematical Geosciences*, *43*(1), 99–120.

Papamitsiou, Z., & Economides, A. A. (2017). Exhibiting achievement behavior during computer-based testing: What temporal trace data and personality traits tell us? *Computers in Human Behavior*, *75*, 423–438.

Paudel, S., Prasad, P. W. C., Alsadoon, A., Islam, M. R., & Elchouemi, A. (2018, July). *Feature selection approach for twitter sentiment analysis and text classification based on Chi-Square and Naïve Bayes*. In International Conference on Applications and Techniques in Cyber Security and Intelligence (pp. 281−298).

Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: language use as an individual difference. *Journal of Personality and Social Psychology*, *77*(6), 1296–1312.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. Austin, TX: University of Texas at Austin. doi: 10.15781/T29G6Z

Pilkonis, P. A. (1977). The behavioral consequences of shyness. *Journal of Personality*, *45*(4), 596–611.

Qiu, L., Lin, H., Ramsay, J., & Yang, F. (2012). You are what you tweet: personality expression and perception on twitter. *Journal of Research in Personality*, *46*(6), 710–718.

Qiu, L., Lu, J. H., Ramsay, J., Yang, S. S., Qu, W. N., & Zhu, T. Z. (2017). Personality expression in Chinese language use. *International Journal of Psychology*, *52*(6), 463–472.

Rubin, K. H., Coplan, R. J., Fox, N. A., & Calkins, S. D. (1995). Emotionality, emotion regulation, and preschoolers' social adaptation. *Development and Psychopathology*, *7*(1), 49–62.

Sandoval, J., & Echandia, A. (1994). Behavior assessment system for children. *Journal of School Psychology*, *32*(4),

419–425.
Sato, E., Matsuda, K. H., & Carducci, B. J. (2018). A factor analytical investigation of the Japanese translation of the Cheek-Buss Shyness Scale in support of the three-component model of shyness. *Personality and Individual Differences*, *124*, 160–167.
Schlenker, B. R., & Leary, M. R. (1982). Social anxiety and self-presentation: A conceptualization model. *Psychological Bulletin*, *92*(3), 641–669.
Shen, T. C., Jia, J., Shen, G. Y., Feng, F. L., He, X. N., Luan, H. B., … Hall, W. (2018, July). *Cross-domain depression detection via harvesting social media*. In Proceedings of the 27th International Joint Conference on Artificial Intelligence (pp. 1611–1617).
Skowron, M., Tkalčič, M., Ferwerda, B., & Schedl, M. (2016, April). *Fusing social media cues: Personality prediction from Twitter and Instagram*. In Proceedings of the 25th International Conference Companion on World Wide Web (pp. 107–108).
Sun, Y. Q., Fan, Y. M., Wang, P., Gong, R. Y., & Gao, F. Q. (2009). Relationship among shyness, posttraumatic stress disorder symptom and mental health of children who lost their relatives in WenChuan earthquake. *Chinese Journal of Clinical Psychology*, *17*(4), 484–486.
[孙源泉, 范一鸣, 王鹏, 宫瑞莹, 高峰强. (2009). 震区丧亲儿童羞怯、创伤后应激障碍症状和心理健康之间的关系. *中国临床心理学杂志*, *17*(4), 484–486.]
Tadesse, M. M., Lin, H. F., Xu, B., & Yang, L. (2018). Personality predictions based on user behavior on the Facebook social media platform. *IEEE Access*, *6*, 61959–61969.
Tandera, T., Hendro, Suhartono, D., Wongso, R., & Prasetio, Y. L. (2017). Personality prediction system from Facebook users. *Procedia Computer Science*, *116*, 604–611.
Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1), 24–54.
Ting, K. M. (2002). An instance-weighting method to induce cost-sensitive trees. *IEEE Transactions on Knowledge and Data Engineering*, *14*(3), 659–665.
Wan, W. Y., Sun, J. M., Liu, J. H., Yang, S. W., Liu, M. M., Xue, J., … Liu, X. Q. (2019). Using social media to explore the linguistic features in female adults with childhood sexual abuse by Linguistic Inquiry and Word Count. *Human Behavior and Emerging Technologies*, *1*(3), 181–189.
Wang, J. Y., Gan, S. Q., Zhao, N., Liu, T. L., Zhu, T. S. (2016). Chinese mood variation analysis based on Sina Weibo. *Journal of University of Chinese Academy of Sciences*, *33*(6), 815–824.
[汪静莹, 甘硕秋, 赵楠, 刘天俐, 朱廷劭. (2016). 基于微博用户的情绪变化分析. *中国科学院大学学报*, *33*(6), 815–824.]
Xu, C. R. (2001). *The Study on the Relationships Among Elementary School Students' Self-esteem, Locus of Control, Parenting Style and Shyness* (Unpublished master's thesis). National Pingtung University.
[徐春蓉. (2001). *国小学童的自尊、内外控、父母管教方式与害羞关系之研究*(硕士学位论文). 国立屏东师范学院.]
Xue, D., Wu, L. F., Hong, Z., Guo, S. Z., Gao, L., Wu, Z. Y., Zhong, X. F., & Sun, J. S. (2018). Deep learning-based personality recognition from text posts of online social networks. *Applied Intelligence*, *48*(11), 4232–4246.
Yarkoni, T. (2010). Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality*, *44*(3), 363–373.
Yuan, C. X., Wu, J. J., Li, H., & Wang, L. H. (2018, July). *Personality recognition based on user generated content*. In 2018 15th International Conference on Service Systems and Service Management (pp. 1–6).
Zhang, C., Han, P. Y., Wang, X. Z., Huang, X. Y., Du, H., Wang, M. J., & Deng, X. (2012). Study on correlation between social anxiety and parental rearing style and shyness of children. *Medical Journal of Chinese People's Health*, *24*(19), 2401–2402.
[张昌, 韩佩英, 王秀珍, 黄晓燕, 都弘, 王美娟, 邓欣. (2012). 儿童羞怯感与社交焦虑和父母养育方式相关性研究. *中国民康医学*, *24*(19), 240–2402.]
Zhao, N., Wang, Y. L., Li, S. J., Liu, X. Q., Wu, P. J., & Zhu, T. S. (2020). Psychological and behavioral impact of Wuhan lockdown and suggestions. *Bulletin of Chinese Academy of Sciences*, *35*(3), 264–272.
[赵楠, 王艺霖, 李思嘉, 刘晓倩, 吴佩璟, 朱廷劭. (2020). 武汉"封城"对民众心理行为的影响与对策建议. *中国科学院院刊*, *35*(3), 264–272.]
Zhu, T. S. (2016). *Psychological research and application in the era of big data*. Beijing: Science Press.
[朱廷劭. (2016). *大数据时代的心理学研究及应用*. 北京: 科学出版社.]
Zimbardo, P. G. (1977). *Shyness: What it is what to do about it*. Boston: Addison-Wesley Publishing Company.
Zimbardo, P. G. (1982). Shyness and the stresses of the human connection. In L. Goldberger & S. Breznitz (Eds.), *Handbook of stress: Theoretical and clinical aspects* (pp. 466–481). New York: Free Press.
Zimbardo, P. G., Pilkonis, P. A., & Norwood, R. M. (1975). The social disease called shyness. *Psychology Today*, *8*(12), 68–72.

Shyness prediction and language style model construction of elementary school students

LUO Fang[1], JIANG Liming[1], TIAN Xuetao[2], XIAO Mengge[1], MA Yanzhen[3], ZHANG Sheng[3]

([1] *School of Psychology, Beijing Normal University, Beijing* 100875, *China*) ([2] *School of Computer and Information Technology, Beijing Jiaotong University, Beijing* 100044, *China*) ([3] *Collaborative Innovation Center of Assessment toward Basic Education Quality, Beijing Normal University, Beijing* 100875, *China*)

Abstract

The present study aimed to explore a new method of measuring shyness based on 1306 elementary school students' online writing texts. A supervised learning method was used to map students' labels (tagged by their results of scale) with their text features (extracted from online writing texts based on a psychological dictionary) to build a machine learning model. Key feature sets for different dimensions of shyness were built and a machine learning model was constructed based on the selected feature to achieve automatic prediction.

The labels were obtained through "National School Children Shyness Scale" completed online by elementary students. The scale includes three dimensions of shyness: shy behavior, shy cognition and shy emotion. Students with Z-scores of each dimension over 1 were labeled as shy and others were labeled as normal. Students' online writing texts were collected from "TeachGrid" (https://www.jiaokee.com/), an online learning platform wherein students writing texts.

The dictionary applied in the present study was Textmind, a widely used Chinese psychological dictionary developed based on Linguistic Inquiry and Word Count (LIWC). The dictionary was compiled mainly based on the corpus of adults. To ensure the validity of extracted features, we modified the original dictionary by expanding the categories and vocabulary with the real writing text of elementary students. The revised dictionary contained 118 categories.

Features were extracted based on the revised dictionary. Chi-square algorithm was applied to identify the features that can distinguish between shy and normal groups to the greatest extent. Three sets of key features confirmed a significant lexical difference between shy and normal individuals. Among the selected features, some were shared by multiple dimensions reflecting the universal textual expression of shy individuals (e.g., The average number of words per sentence and the frequency of social words of shy individuals were less than that of normal counterparts.), and there were certain features reflected the unique characteristics of certain dimension (Perception words predicted shy behavior reflecting that high shy behavior individuals frequently felt being watched).

Based on the selected features, Python 3.6.2 was used to construct the six prediction modes: Decision Tree, Random Forest, Support Vector Machine, Logistic Stitch Regression, K-Nearest Neighbor and Multilayer Perceptron. Overall, random forests have achieved the best results in the present study. The F1 score was 0.582, 0.552 and 0.545 for behavior cognition and emotion showing the feasibility of automatically predicting shyness characteristics of elementary school students based on textual language. The implication of word embedding, and deep learning models would improve the final prediction.

Key words shyness, online writing, psychological dictionary, text mining, language style model